STATS 200: Introduction to Statistical Inference — Autumn 2016

Lecture 17 — Plugin estimators and the delta method

17.1 Estimating a function of θ

In the setting of a parametric model, we have been discussing how to estimate the parameter θ. We showed how to compute the MLE $\hat{\theta}$, derived its variance and sampling distribution for large n, and showed that no unbiased estimator can achieve variance much smaller than that of the MLE for large n (the Cramer-Rao lower bound).

In many examples, the quantity we are interested in is not θ itself, but some value $g(\theta)$. The obvious way to estimate $g(\theta)$ is to use $g(\hat{\theta})$, where $\hat{\theta}$ is an estimate (say, the MLE) of θ. This is called the **plugin estimate** of $g(\theta)$, because we are just "plugging in" $\hat{\theta}$ for θ.

Example 17.1 (Odds)**.** You play a game with a friend, where you flip a biased coin. If the coin lands heads, you give your friend \$1. If the coin lands tails, your friend gives you \$$x$. What is the value of x that makes this a fair game?

If the coin lands heads with probability p, then your expected winnings is $-p+(1-p)x$. The game is fair when $-p+(1-p)x=0$, i.e. when $x=p/(1-p)$. This value $p/(1-p)$ is the *odds* of getting heads to getting tails. To estimate the odds from n coin flips

$$X_1,\ldots,X_n \overset{IID}{\sim} \text{Bernoulli}(p),$$

we may first estimate p by $\hat{p}=\bar{X}$. (This is both the method of moments estimator and the MLE.) Then the plugin estimate of $p/(1-p)$ is simply $\bar{X}/(1-\bar{X})$.

The odds falls in the interval $(0,\infty)$ and is not symmetric about $p=1/2$. We oftentimes think instead in terms of the log-odds, $\log\frac{p}{1-p}$—this can be any real number and is symmetric about $p=1/2$. The plugin estimate for the log-odds is $\log\frac{\bar{X}}{1-\bar{X}}$.

Example 17.2 (The Pareto mean)**.** The Pareto(x_0,θ) distribution for $x_0>0$ and $\theta>1$ is a continuous distribution over the interval $[x_0,\infty)$, given by the PDF

$$f(x|x_0,\theta)=\begin{cases}\theta x_0^\theta x^{-\theta-1} & x\geq x_0\\ 0 & x<x_0.\end{cases}$$

It is commonly used in economics as a model for the distribution of income. x_0 represents the minimum possible income; let's assume that x_0 is known and equal to 1. We then have a one-parameter model with PDFs $f(x|\theta)=\theta x^{-\theta-1}$ supported on $[1,\infty)$.

The mean of the Pareto distribution is

$$\mathbb{E}_\theta[X]=\int_1^\infty x\cdot\theta x^{-\theta-1}dx=\theta\frac{x^{-\theta+1}}{-\theta+1}\bigg|_1^\infty=\frac{\theta}{\theta-1},$$

so we might estimate the mean income by $\hat{\theta}/(\hat{\theta}-1)$ where $\hat{\theta}$ is the MLE. To compute $\hat{\theta}$ from observations $X_1, \ldots, X_n$, the log-likelihood is

$$l(\theta) = \sum_{i=1}^{n} \log(\theta {X_i}^{-\theta-1}) = \sum_{i=1}^{n} (\log\theta - (\theta+1)\log X_i) = n\log\theta - (\theta+1)\sum_{i=1}^{n}\log X_i.$$

Solving the equation

$$0 = l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \log X_i$$

yields the MLE $\hat{\theta} = n/\sum_{i=1}^{n} \log X_i$.

17.2 The delta method

We would like to be able to quantify our uncertainty about $g(\hat{\theta})$ using what we know about the uncertainty of $\hat{\theta}$ itself. When n is large, this may be done using a first-order Taylor approximation of g, formalized as the **delta method**:

Theorem 17.3 (Delta method)**.** *If a function $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ_0 with $g'(\theta_0) \neq 0$, and if*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, v(\theta_0))$$

in distribution as $n \to \infty$ for some variance $v(\theta_0)$, then

$$\sqrt{n}(g(\hat{\theta}) - g(\theta_0)) \to \mathcal{N}(0, (g'(\theta_0))^2 v(\theta_0))$$

in distribution as $n \to \infty$.

Proof sketch. We perform a Taylor expansion of $g(\hat{\theta})$ around $\hat{\theta} = \theta_0$:

$$g(\hat{\theta}) \approx g(\theta_0) + (\hat{\theta} - \theta_0) g'(\theta_0).$$

Rearranging yields

$$\sqrt{n}\left(g(\hat{\theta}) - g(\theta_0)\right) \approx \sqrt{n}(\hat{\theta} - \theta_0) g'(\theta_0),$$

and multiplying a mean-zero normal variable by a constant c scales its variance by c^2. □

Example 17.4 (Log-odds)**.** Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$, and recall the plugin estimate of the log-odds $\log \frac{p}{1-p}$ given by $\log \frac{\bar{X}}{1-\bar{X}}$. By the Central Limit Theorem,

$$\sqrt{n}(\bar{X} - p) \to \mathcal{N}(0, p(1-p))$$

in distribution, where $p(1-p)$ is the variance of a Bernoulli(p) random variable. The function $g(p) = \log \frac{p}{1-p} = \log p - \log(1-p)$ has derivative

$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)},$$

so by the delta method,

$$\sqrt{n}\left(\log\frac{\bar{X}}{1-\bar{X}} - \log\frac{p}{1-p}\right) \to \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).$$

In other words, our estimate of the log-odds of heads to tails is approximately normally distributed around the true log-odds $\log\frac{p}{1-p}$, with variance $\frac{1}{np(1-p)}$.

Suppose we toss this biased coin $n = 100$ times and observe 60 heads, i.e. $\bar{X} = 0.6$. We would estimate the log-odds by $\log\frac{\bar{X}}{1-\bar{X}} \approx 0.41$, and we may estimate our standard error by $\sqrt{\frac{1}{n\bar{X}(1-\bar{X})}} \approx 0.20$.

Example 17.5 (The Pareto mean)**.** Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(1, \theta)$, and recall that the MLE for θ is $\hat{\theta} = n/\sum_{i=1}^n \log X_i$. We may use the maximum-likelihood theory developed in Lecture 14 to understand the distribution of $\hat{\theta}$: We compute (for $x \geq 1$)

$$\begin{aligned}
\log f(x|\theta) &= \log(\theta x^{-\theta-1}) = \log\theta - (\theta+1)\log x \\
\frac{\partial}{\partial\theta}\log f(x|\theta) &= \frac{1}{\theta} - \log x \\
\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) &= -\frac{1}{\theta^2}.
\end{aligned}$$

Then the Fisher information is given by $I(\theta) = 1/\theta^2$, so

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \theta^2)$$

in distribution as $n \to \infty$. For the function $g(\theta) = \theta/(\theta-1)$, we have

$$g'(\theta) = \frac{1}{\theta-1} - \frac{\theta}{(\theta-1)^2} = -\frac{1}{(\theta-1)^2}.$$

So the delta method implies

$$\sqrt{n}\left(\frac{\hat{\theta}}{\hat{\theta}-1} - \frac{\theta}{\theta-1}\right) \to \mathcal{N}\left(0, \frac{\theta^2}{(\theta-1)^4}\right).$$

Say, for a data set with $n = 1000$ income values, we obtain the MLE $\hat{\theta} = 1.5$. We might then estimate the mean income as $\hat{\theta}/(\hat{\theta}-1) = 3$, and estimate our standard error by $\sqrt{\frac{\hat{\theta}^2}{n(\hat{\theta}-1)^4}} \approx 0.19$.

What if we decided to just estimate the mean income by the sample mean, $\bar{X}$? Since $\mathbb{E}[X_i] = \theta/(\theta-1)$, the Central Limit Theorem implies

$$\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta-1}\right) \to \mathcal{N}(0, \text{Var}[X_i])$$

in distribution. For $\theta > 2$, we may compute

$$\mathbb{E}[X_i^2] = \int_1^\infty x^2 \cdot \theta x^{-\theta-1}dx = \theta\frac{x^{-\theta+2}}{-\theta+2}\bigg|_1^\infty = \frac{\theta}{\theta-2},$$

so

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{\theta}{\theta-2} - \left(\frac{\theta}{\theta-1}\right)^2 = \frac{\theta}{(\theta-1)^2(\theta-2)}.$$

(If $\theta \leq 2$, the variance of X_i is actually infinite.) For any θ, this variance is greater than $\theta^2/(\theta-1)^4$.

Thus if the Pareto model for income is correct, then our previous estimate $\hat{\theta}/(\hat{\theta}-1)$ is more accurate for the mean income than is the sample mean $\bar{X}$. Intuitively, this is because the Pareto distribution is heavy-tailed, and the sample mean $\bar{X}$ is heavily influenced by rare but extremely large data values. On the other hand, $\hat{\theta}$ is estimating the shape of the Pareto distribution and estimating the mean by its relationship to this shape in the Pareto model. The formula for $\hat{\theta}$ involves the values $\log X_i$ rather than X_i, so $\hat{\theta}$ is not as heavily influenced by extremely large data values. Of course, the estimate $\hat{\theta}/(\hat{\theta}-1)$ relies strongly on the correctness of the Pareto model, whereas $\bar{X}$ would be a valid estimate of the mean even if the Pareto model doesn't hold true.

That the plugin estimate $g(\hat{\theta})$ performs better than $\bar{X}$ in the previous example is not a coincidence—it is in certain senses the best we can do for estimating $g(\theta)$. For example, we have the following more general version of the Cramer-Rao lower bound:

Theorem 17.6. *For a parametric model $\{f(x|\theta) : \theta \in \Omega\}$ (satisfying certain mild regularity assumptions) where θ is a single parameter, let g be any function differentiable on all of Ω, and let T be any unbiased estimator of $g(\theta)$ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$. Then*

$$\mathrm{Var}_\theta[T] \geq \frac{g'(\theta)^2}{nI(\theta)}.$$

The proof is identical to that of Theorem 15.4, except with the equation $\theta = \mathbb{E}_\theta[T]$ replaced by $g(\theta) = \mathbb{E}_\theta[T]$. (Differentiating this equation yields $g'(\theta) = \mathbb{E}_\theta[TZ] = \mathrm{Cov}_\theta[T, Z]$ as in Theorem 15.4.) An estimator T for $g(\theta)$ that achieves this variance $g'(\theta)^2/(nI(\theta))$ is **efficient**. The plugin estimate $g(\hat{\theta})$ where $\hat{\theta}$ is the MLE achieves this variance asymptotically, so we say it is **asymptotically efficient**. This theorem ensures that no unbiased estimator of $g(\theta)$ can achieve variance much smaller than $g(\hat{\theta})$, when n is large, and in particular applies to the estimator $T = \bar{X}$ of the previous example.